第53回(平成８年後期)全国大会

講演論文集(3)

データベースとメディア

- データベース
- 情報検索
- メディアと情報

ネットワークサイエンス

- マルチメディア通信と分散処理
- 分散システム運用
- マルチメディア符号化

グローバル・ネットワーク

平成８年９月４日～６日　　於：大阪工業大学

社団法人 情報処理学会
Information Processing Society of Japan

WWWサーバとデータベースの連携システムの利用状況分析ツールの開発

4T－4

彭 智勇　星野 寛
京都高度技術研究所

1　まえがき

インターネットの普及に伴い、WWWサーバによって発信する情報が膨大になっている。その中には、頻繁に更新する必要のあるものが多い。WWWサーバの膨大な情報がデータベースによって管理されれば、その検索、更新を効率よくすることができる。WWWサーバとデータベースの連携によって情報サービスを提供する際に、インタネットの上で、だれでもアクセスして来られるため、その利用状況の把握が情報サービスの運用に対して重要だと考えられる。既存のWWWサーバの利用状況分析ツール(wwwstat[1], getstats[2]など)は、データベースから動的に生成されたページの集計機能をサポートしていない。特に、ユーザの特徴データ(年齢、性別、職業、国籍など)を考慮していないので、情報サービスの利用状況を細かく分析することができない。ユーザの特徴データをアクセスの集計に使うのはアクセスしてくるユーザの認識を前提としている。しかし、ユーザとデータベースの間に介されているWWWサーバがステートレスプロトコル(HTTP)を用いているため、ユーザの認識が困難になる。本稿では、WWWサーバとデータベースの連携システムに、ユーザの認識機構を新たに導入し、その利用状況分析ツールを提案する。本ツールでは、ユーザの特徴データを考慮した上で、アクセスの履歴データから、データベースのビューによって、種々の利用状況分析のための集計データを発掘することができる。

2　ユーザの特徴データの登録

本ツールは、ユーザが使っているブラウザーに、ユーザ名、パスワード、氏名、性別、年齢、職業、国籍、住所などを入力できるようなフォームを出して、ユーザはフォームに応じてその特徴データを入力する。ユーザは入力フォームの登録ボタンをクリックすると、その特徴データをWWWサーバに送ることになる。本ツールはユーザに唯一の識別子を発行して、それを送られたユーザの特徴データと一緒に、データベースのユーザ特徴データ表に登録する。その次、すでに登録したユーザはシステムのホームページにユーザ名とパスワードを入力するだけで、システムを利用することができる。

Development of Access Statistics Analysis Tools for an Integrated System of WWW Server and Database
Zhiyong Peng, Hiroshi Hoshino
Advanced Software Technology & Mechatronics Research Institute of Kyoto

3　ユーザの認識機構の導入

通常、データの検索は複数のアクセスからなる。例えば、ユーザが、まず、検索条件を入力すると、検索結果のリストを返してくれる。次に、リストを見て、項目を選択すると、その詳細情報が表示されることになる。しかし、ユーザが使っているブラウザーとWWWサーバの交信はステートレスプロトコル(HTTP)に基づいて行なわれる。つまり、アクセスの要求があるたびに接続して、1ページ分のデータを取得するとその時点で接続が切れる。アクセスしてきたユーザを認識する方法として、ユーザ名とパスワードをアクセスするたびに提供させるのはユーザにとって非常に不便である。この問題は図1に示すようにユーザの識別子をユーザブラウザーとWWWサーバの間に伝達させることで解決することができる。

ユーザの認識に使われるユーザの識別子はシステムのホームページに入力されたユーザ名とパスワードを使って、データベースに登録したユーザ特徴データ表から得られる。連携システムでは、ホームページ以外のページが動的に生成されるため、ユーザの識別子をページに書き込んで、ブラウザーに送ることができる。例えば、検索条件入力ページを生成するとき、ユーザの識別子を隠れインプットフィールドとして以下のようにこのページに書き込むと、それをブラウザー側には表示せず、入力した検索条件と一緒にWWWサーバへ渡すことができる。

```
< FORM METHOD="post" ACTION="/cgi-bin/dbgate/query" >
< INPUT NAME="userID" TYPE="hidden" VALUE="A" >
< 検索条件の入力部分 >
< INPUT TYPE="submit" > < INPUT TYPE="reset" >
< /FORM >
```

それによって渡されたユーザの識別子を使って、利用状況分析ツールは「どのユーザからのアクセスであるか」判断することができる。また、ユーザの識別子はページ中のリンクにも隠すことができる。例えば、検索結果リストページを生成するとき、ユーザの識別子を次のようにこのページに入れれば、リンクボタンをクリックすると、ユーザの識別子を再びWWWサーバに渡すことができる。

```
< A HREF="/cgi-bin/dbgate/getinfo?infoID=00014&userID=A" > </A >
```

つまり、ブラウザーとWWWサーバの交信中に、ユーザの識別子の伝達を上記の方式で繰り返して、ユーザ名とパスワードを入力することなく、アクセスしたユーザを認

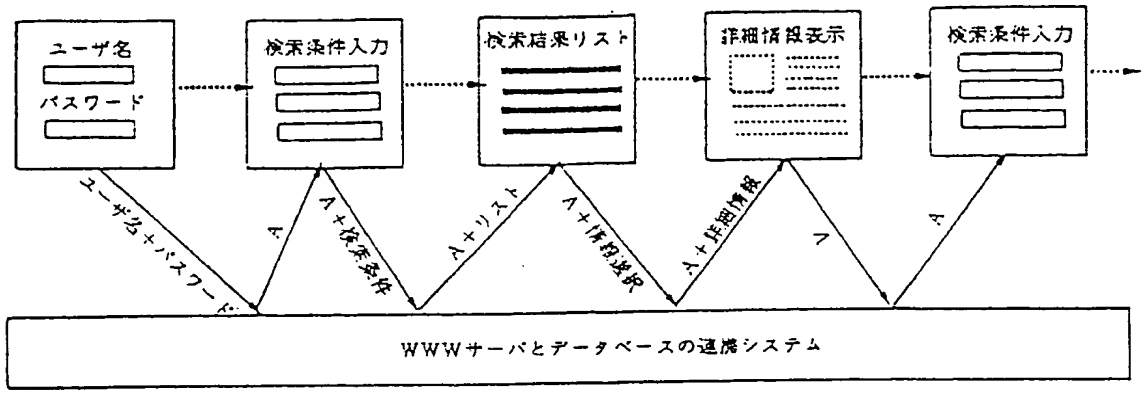

図 1: ユーザ識別子の伝達によるアクセスしてきたユーザの認識

識できるようにする。

4 アクセスの履歴データの収集

WWW サーバとデータベースの連携のためのゲートウェイは WWW サーバの外部プログラムとして実現される。WWW サーバは、ブラウザーからの入力を外部プログラムに渡して処理を行ない、その結果に基づいて、レスポンスを返すための CGI(Common Gateware Interface) を提供している。ゲートウェイは、CGI によって、ユーザからの検索条件を使って、データベースの SQL コマンドを呼び出し、情報検索を行なう。最後に、検索結果を HTML ページに変換して、レスポンスとして返す。

ゲートウェイはユーザからのアクセスの要求に従ってページを動的に生成すると同時に、アクセスの履歴データを収集することもできる。その時点で、ページの生成中に使われているデータとそれらの間の関連が分かるので、データ単位で (WWW サーバの集計ツールはページ単位で)、必要なアクセスの履歴データのみを記録することが可能になる。本ツールは、集計のデータ単位を、データベースの中で識別できるようなデータ組とする。例えば、関係データベースの表の一行やオブジェクト指向データベースの一つのオブジェクトなどが、集計のデータ単位として扱われる。そのデータ組がそれぞれの識別子 (キー、オブジェクト識別子など) によって区別されるので、アクセスの履歴データをユーザの識別子、アクセスしたデータの識別子、日付および処理時刻の組とする。

5 アクセスの集計とその結果の活用

上記のように、ユーザの特徴データとアクセスの履歴データがデータベースの表として貯蓄されている。これらの表から、様々なビューを導出することによって、種々の利用状況分析のためのアクセスの集計を行なえる。

アクセスの集計結果が情報サービスの運用に役立つと考えられる。例えば、データの各時間帯のアクセス頻度が分かれば、データの利用時間傾向を把握できるため、よくアクセスされるデータは時間に応じて動的にデータベースシステムの側でキャッシュさせるように、情報の発信を効率よくすることができる。また、商品広告サービスを提供する場合に、年齢別のアクセス頻度表を導出することによって、各年齢ごとの人気商品も掘り出すことができる。その他にも、ユーザがアクセスした商品データのセットを抽出することができるため、人気な商品セットのパターンを発掘することも可能である。それに従って、商品データをセットの形で発信すると、ユーザのアクセスの命中率を高めることになる。

6 むすび

利用状況分析ツールを実現するために、まだ、いろいろな問題点を検討しなければならない。例えば、動的に生成されるページがブラウザーまたは proxy サーバでキャッシュされれば、データベースの最新状態を正確に反映できない可能性があるし、アクセス回数に誤差を生じることもあり得る。また、ユーザの識別子を含めたページが他のユーザに公開されると、アクセス回数の誤差がさらに広がる。Netscape の Cookie という新しい機能を利用して、これらの問題を解決することが可能である。そのほか、動的なページの生成に使われる検索結果には、その検索に役立つデータの識別子が含まれていない場合があるため、データベースが検索を行いながら、アクセスの集計に必要なデータ識別子を収集する機構の開発も必要である。

謝辞: ご貴重なコメントを頂いた京都高度技術研究所の新井様に感謝致します。

参考文献

[1] wwwstat:http://www.ics.uci.edu/WebSoft/wwwstat/

[2] getstats:http://www.eit.com/software/getstats/getstats.html